UNIVERSITY OF
BIRMINGHAM

Adventures in storage

Simon Thompson

Research Computing Infrastructure Architect

IT Services

University of Birmingham

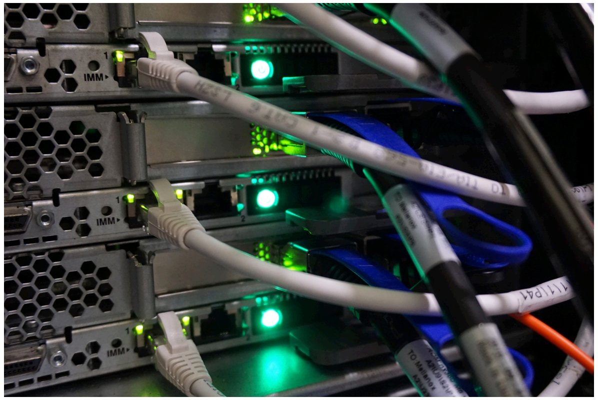
- Royal charter in 1900 (history back to 1828)
- Member of Russell Group (Research Intensive Universities)
- 33,351 students (2016/17)
- 11 staff and alumni are Nobel prize winners
- £135M research income (2015/16)
- Dubai campus opens 2018

Powering research …

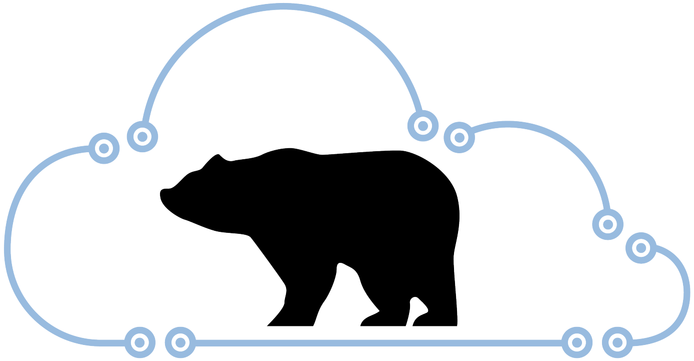
BEAR

BIRMINGHAM ENVIRONMENT
FOR ACADEMIC RESEARCH

Powering research …



BlueBEAR HPC



BEAR Cloud

Research Data Storage (>PB)

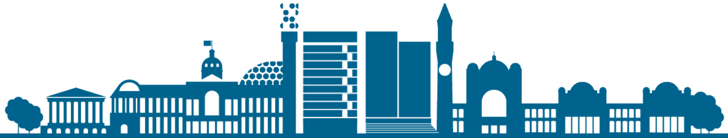
Research Data Archive

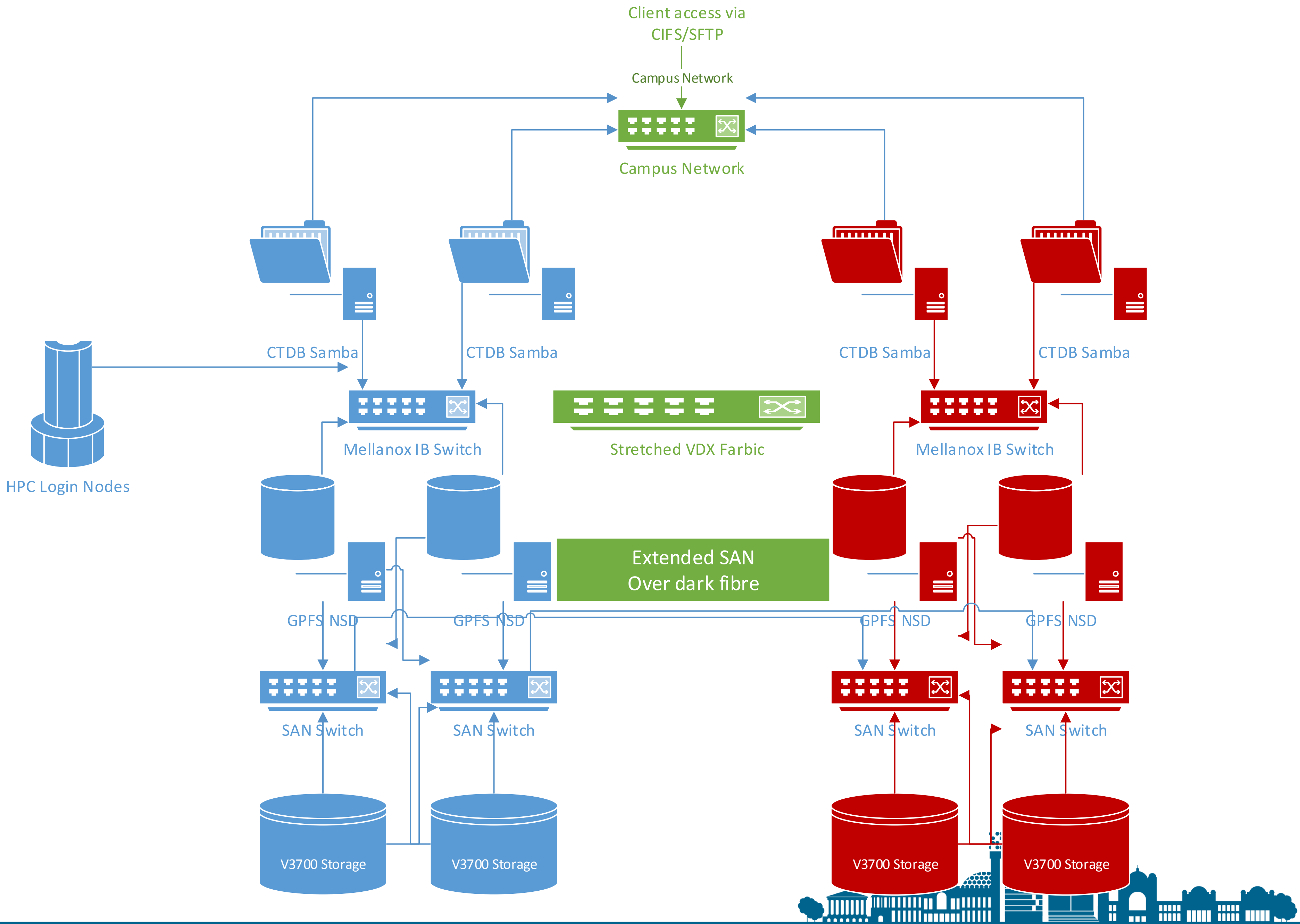
Research Network

Our storage

□ Research Data Store & Archive

- IBM Spectrum Scale™ (Data management edition)
- Sync replication between data centres
- Runs on Storwize®, DDN® SFA
- IBM Spectrum Protect™ with ILM for tape/SOBAR





Mellanox and Switch-IB are trademarks of Mellanox Technologies.

Why did we want AFM?

- Bulk data store is copies=2
- Slows HPC workload waiting for write
- Large numbers of HPC nodes in cluster all need to talk to bulk data store
  - Issues = denial of service ;-)
- Building new DCs with no IB links
  - How can we maintain service with no access to bulk data store?

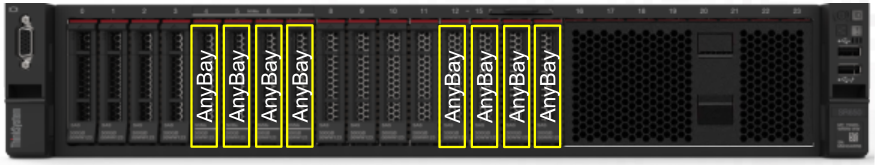
Lenovo DSS-G Roadmap Outlook

- GSS 3.2 (Dec/2017)
  - GSS is in „maintenance mode“ now, hardware will be withdrawn from marketing Dec/2017
  - Software: RHEL 7.3; latest GPFS 4.2.3 and 4.1.1 PTFs; MOFED 4.2; OPA 10.6; xCAT 2.13
  - Further GSS releases in 2018 ... incl. RHEL 7.4 support in 2H2018

- DSS-G 1.2 (Dec/2017)
  - Rackless DSS-G (order without a 1410 rack)
  - 12TB NL-SAS drives
  - Software: RHEL 7.3; latest GPFS 4.2.3 and 4.1.1 PTFs; MOFED 4.2; OPA 10.6; xCAT 2.13
  - DSS G100 NVMe server: 1-8 NVMe drives; classical Spectrum Scale (not GNR)
- DSS-G 2.0 (Mar/2018)
  - Server transition to Lenovo SR650 (Purley)
  - 4x SAS adapters (Lenovo 430-16e)
  - Latest software levels; disk drive refreshes

Lenovo DSS G100 NVMe Server (Dec/2017)

- Lenovo ThinkSystem SR650 server
  - 2x 6142 SkyLake CPUs; 192/384 GB
  - Up to 8x U.2 NVMe drives in AnyBay slots
  - Networking options:
    - 2x Mellanox ConnectX-5 2-port (VPI)
    - 2x Intel OPA100 1-port
    - Ethernet options: 10 / 25 / 40 / 100GbE



- Software stack: „Classical“ Spectrum Scale
  - GPFS replication as needed for redundancy

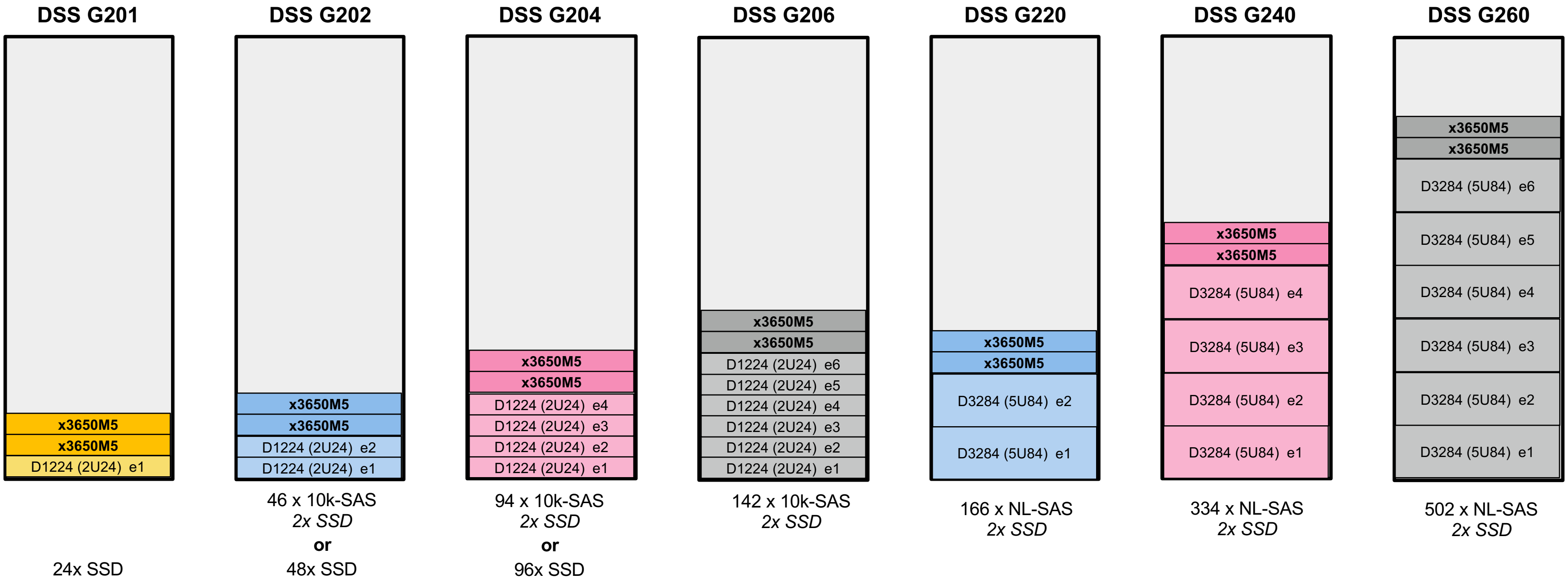
- Initial bandwidth testing on 2x G100 nodes:
  - Up to 34 GB/s read; 30 GB/s write (GPFS 4.2.3)

NVMe drive support at GA
(adding Intel P4600 and P4500 slightly later in Dec/2017):

NVMe drive model	Capacity	read BW	write BW	read IOPS	write IOPS
	[TB]	[GB/s]	[GB/s]	[1000/s]	[1000/s]
2.5-inch hot-swap SSDs - Performance U.2 NVMe PCIe*					
ThinkSystem U.2 PX04PMB 800GB Performance 2.5" NVMe PCIe 3.0 x4 HS SSD	0,80	3,10	2,35	660	185
ThinkSystem U.2 PX04PMB 1.6TB Performance 2.5" NVMe PCIe 3.0 x4 HS SSD	1,60	3,10	2,35	660	185
2.5-inch hot-swap SSDs - Mainstream U.2 NVMe PCIe*					
ThinkSystem U.2 PX04PMB 960GB Mainstream 2.5" NVMe PCIe 3.0 x4 HS SSD	**0,96**	**3,10**	**2,35**	**660**	**165**
ThinkSystem U.2 PX04PMB 1.92TB Mainstream 2.5" NVMe PCIe 3.0 x4 HS SSD	1,92	3,10	2,35	660	165
2.5-inch hot-swap SSDs - Entry U.2 NVMe PCIe*					
ThinkSystem U.2 PM963 1.92TB Entry 2.5" NVMe PCIe 3.0 x4 Hot Swap SSD	1,92	2,00	1,20	430	40
ThinkSystem U.2 PM963 3.84TB Entry 2.5" NVMe PCIe 3.0 x4 Hot Swap SSD	3,84	2,00	1,20	430	40

Current DSS-G Building Blocks

Distributed Storage Solution for IBM Spectrum Scale™



DSS G201

x3650M5
x3650M5
D1224 (2U24) e1

24x SSD

DSS G202

x3650M5
x3650M5
D1224 (2U24) e2
D1224 (2U24) e1

46 x 10k-SAS
2x SSD
or
48x SSD

DSS G204

x3650M5
x3650M5
D1224 (2U24) e4
D1224 (2U24) e3
D1224 (2U24) e2
D1224 (2U24) e1

94 x 10k-SAS
2x SSD
or
96x SSD

DSS G206

x3650M5
x3650M5
D1224 (2U24) e6
D1224 (2U24) e5
D1224 (2U24) e4
D1224 (2U24) e3
D1224 (2U24) e2
D1224 (2U24) e1

142 x 10k-SAS
2x SSD

DSS G220

x3650M5
x3650M5
D3284 (5U84) e2
D3284 (5U84) e1

166 x NL-SAS
2x SSD

DSS G240

x3650M5
x3650M5
D3284 (5U84) e4
D3284 (5U84) e3
D3284 (5U84) e2
D3284 (5U84) e1

334 x NL-SAS
2x SSD

DSS G260

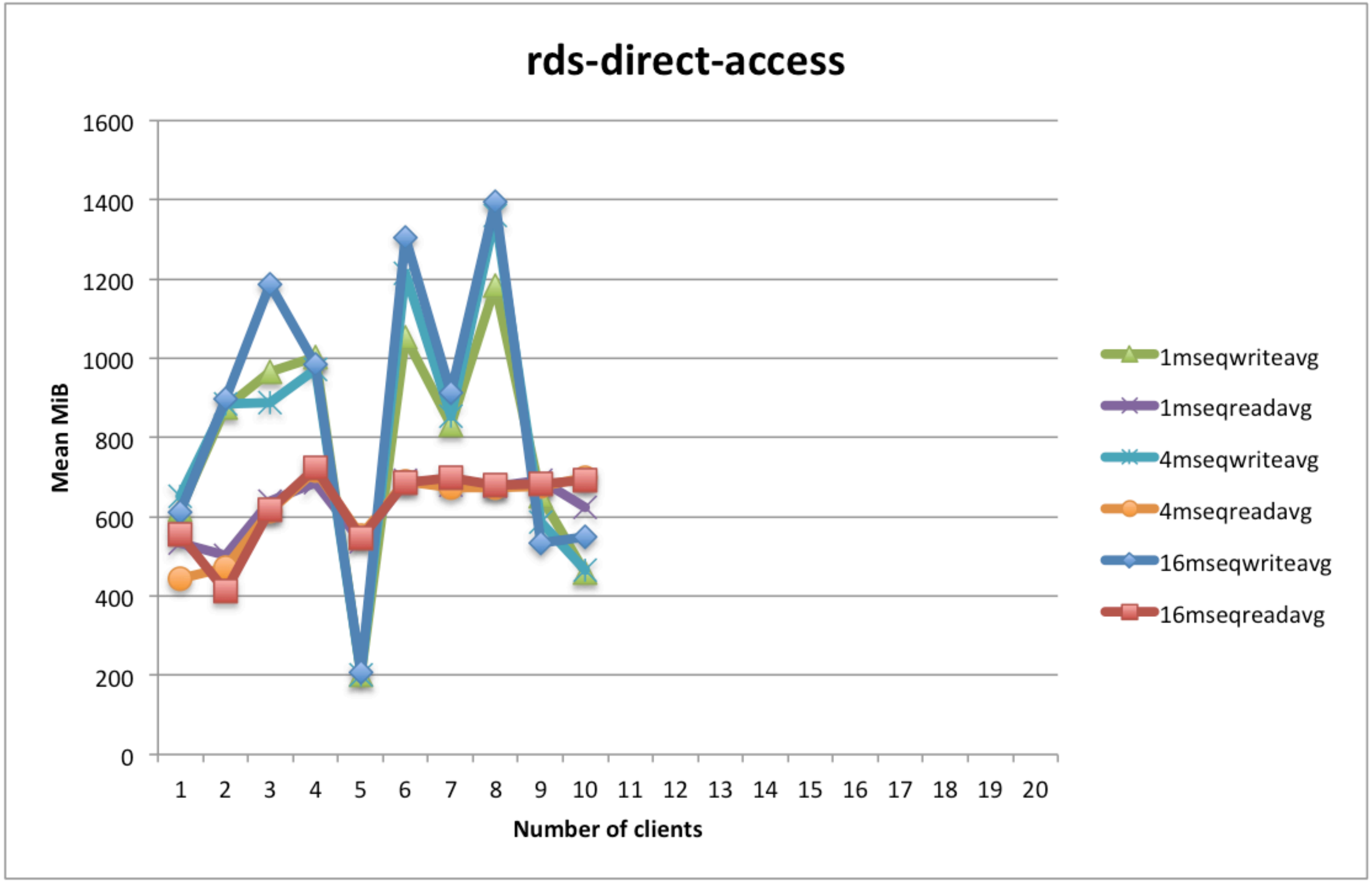
x3650M5
x3650M5
D3284 (5U84) e6
D3284 (5U84) e5
D3284 (5U84) e4
D3284 (5U84) e3
D3284 (5U84) e2
D3284 (5U84) e1

502 x NL-SAS
2x SSD

- Additional building blocks (G210, G230, ...) will be released as soon as supported by IBM Spectrum Scale.

Pretend I am Sven…

- All benchmarks were run at the University of Birmingham
- Results in a different environment may vary
- HPC compute nodes are attached using Mellanox ConnectX-4 EDR + 10GbE
- Benchmarks run with IOR

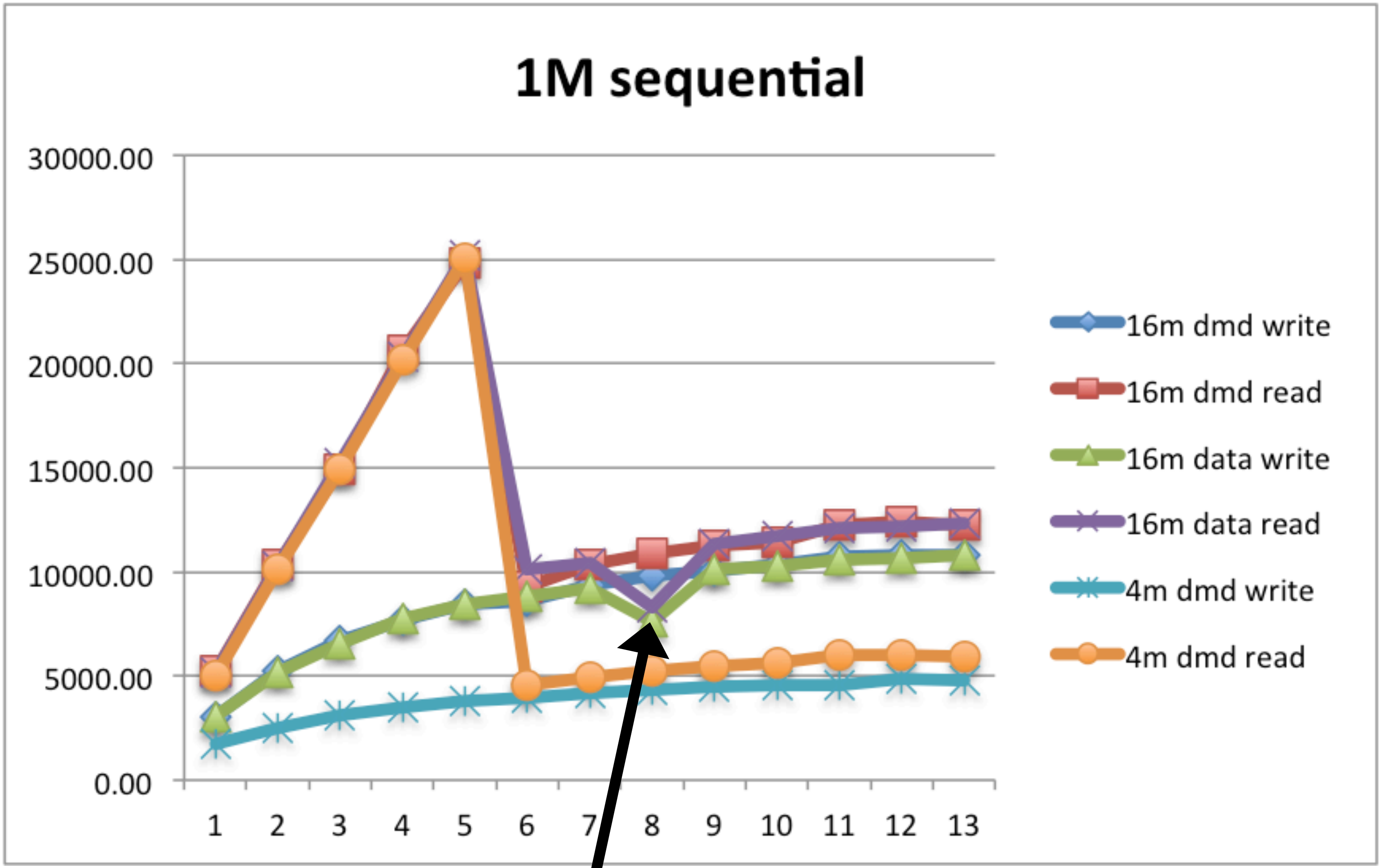

rds-direct-access
Mean MiB
1600
1400
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Number of clients
1mseqwriteavg
1mseqreadavg
4mseqwriteavg
4mseqreadavg
16mseqwriteavg
16mseqreadavg

The Lenovo™ DSS-G …

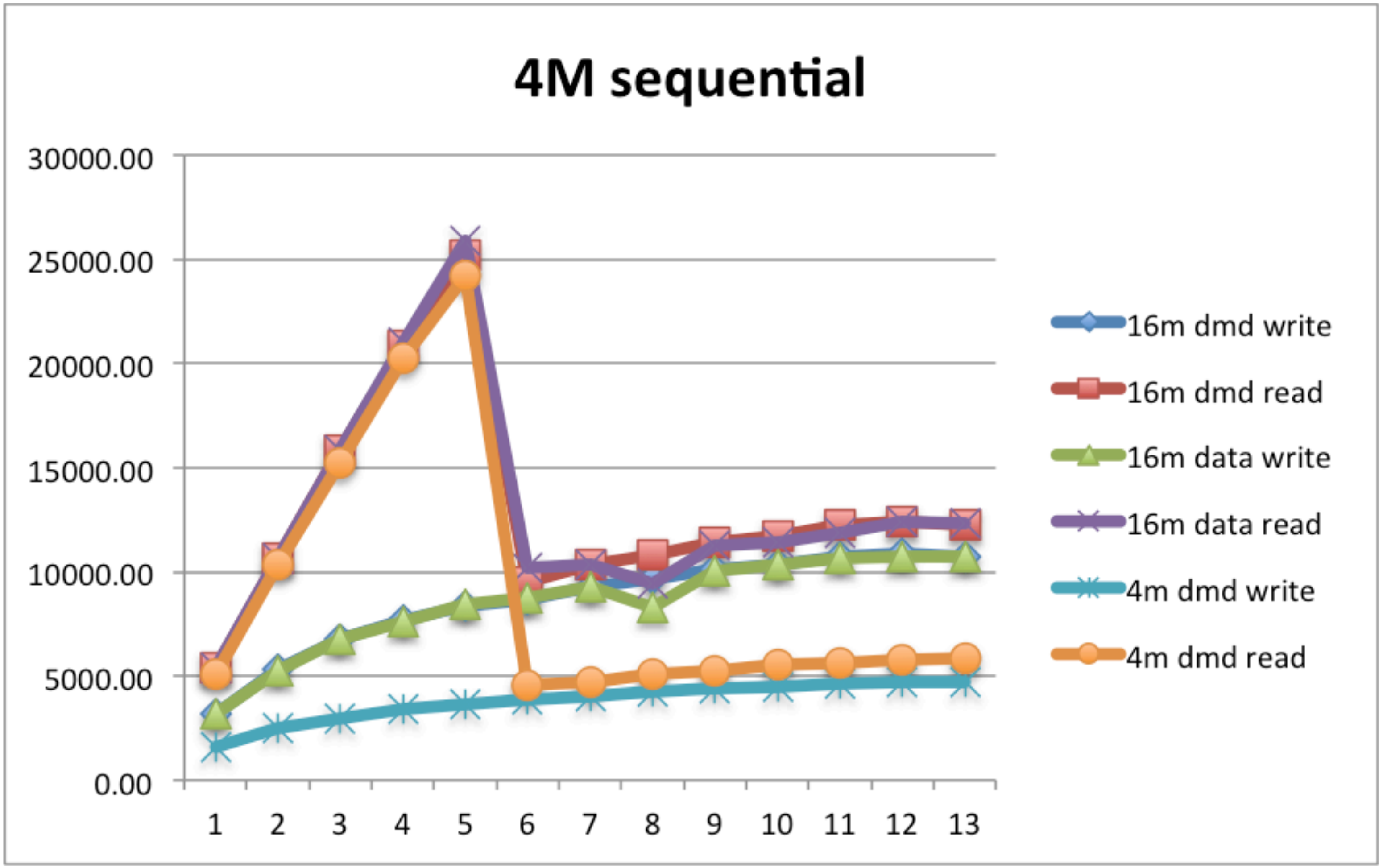
- We learnt:
  - IMM firmware image was broken with VLAN tags
  - GPFS bug when using readReplicaPolicy (fixed)
  - Can’t SOBAR restore an FS with different block size (won’t be fixed ☹)

The Lenovo™ DSS-G …

- □ We also learnt:
  - – Easy to deploy
  - – Easy to upgrade
  - – Can add more storage in a new DA
  - – Mixed data and metadata vs data only almost NO difference
  - – Fun with caches
  - – Drive rebuild = minor impact



Failed drive rebuild


4M sequential
30000.00
25000.00
20000.00
15000.00
10000.00
5000.00
0.00
1
2
3
4
5
6
7
8
9
10
11
12
13
16m dmd write
16m dmd read
16m data write
16m data read
4m dmd write
4m dmd read

And then things broke …

□ Pre-allocate is not supported

- ld.gold uses pre-allocate when linking files
- It was the default linker in our (easy)build
- Workaround with linker flags

□ Truncate was broken

- Its fixed now
- This broke ABAQUS HPC jobs

And then it broke again …

- Huge queues (millions of items)
- AFM gateway nodes randomly crashing

- Working:
  - New files
  - Open existing files
- Broken:
  - Write to existing file

It wasn’t AFM

- SMB users reporting WEIRD ACL issues
  - Copying the files “fixed”
- We had a DMAPI problem at HOME
- With help of IBM Support:
  - AFM side traces
  - Identified error write code
  - Corresponded to a DMAPI issue
- At home DMAPI had failed over to a node that didn’t work with that FS

Stuff I’m not sure about …

- Quotas?
- Home: independent file-sets for years (2015, 2016)
- Home: file-sets per project
- Cache: we have 1 file-set to map to IFS
  - Can I use AFM to migrate to new block size?
- Can you make a node in an ILM HA environment have FS affinity?

Thanks!

BEAR

BIRMINGHAM ENVIRONMENT
FOR ACADEMIC RESEARCH

Email: s.j.thompson@bham.ac.uk